Author's response to reviews

Title: Temporal Stability of Objective Structured Clinical Exams: A Longitudinal Study Employing Item Response Theory

Authors:

Lubna A Baig (lbaig@ucalgary.ca)
Claudio Violato (violato@ucalgary.ca)

Version: 2 Date: 19 November 2012

Author's response to reviews: see over
Dear Dr Danette McKinley

We have revised our manuscript based on your and the reviewers comments. I hope the paper will now be accepted for publication in your journal.

A point by point response to your comments and reviewers comments is given below.

Editor's Comment: (With responses)

This was an interesting paper describing an approach to ensuring the equivalence of OSCE stations when administered to different candidates, with different examiners and SPs over time. While there is not much work done in this area, the authors should consider the work of McKinley and Boulet on drift in case difficulty over time, attributable to the SP [McKinley, D. W., & Boulet, J. R. (2004). Detecting score drift in a high-stakes performance-based assessment. Advances in Health Sciences Education: Theory and Practice, 9(1), 29-38.]

Thank you for bringing this paper to our attention. We have included it and revised our literature review accordingly (pgs. 3, 4, 5-6, 7).

I agreed with the reviewer that the introduction should provide more of a basis for why temporal stability is a concern. Is this a quality assurance measure? How might temporal stability affect the validity of the interpretations associated with this examination.

We have now discussed this more full and in more detail in the introduction (pgs. 6, 7).

The authors should consider explaining more about the administration of the examination, particularly since this information would be relevant to the design employed for the GENOVA. If the SPs do not score the candidates, and only the examiners do, then nesting SP within examiner is reasonable (i.e., these factors are confounded). However, there should be a variance component associated with this nested factor, in addition to a factor for candidates.

More detail is now provided on the administration and details of the GENOVA design and analyses (pgs. 9, 10).

There was not a clear reason presented for the exploratory factor analysis. If the checklists already have a defined structure (i.e., history-taking, physical examination, communication) what were the six factors identified for time 1 and the seven for time 2? Since only 6 stations were in common, how do you interpret the results
for time 2? Are cases loading to factors in the first administration but not the second? How does this relate to temporal stability?

Upon closer reflection, we agree with the reviewers and the editor that EFA was not necessary and indeed only contributes confusion to the analyses. Therefore we have removed it from the paper (pgs. 11, 12-13, 20).

In considering these factors, as well as those presented by the reviewers, ensure that the major compulsory revisions are attended to before resubmitting the paper.

We have responded to all of the reviewer’s comments and indicated this in yellow highlight on their comments pages.

Response to the comments of reviewer one: William Roberts

Major Compulsory Revisions

This review includes a list of comments to address areas of strength and weakness with recommended corrections. Technical and general areas of weakness in the manuscript will need to be addressed before acceptance for publication.

1. Specify in design notation the effects for the single facet nested G-study design. Include if this is a norm referenced or criterion based examination. If this is a criterion based examination, then the index of dependability is the appropriate measure of reliability.

This is a single facet, nested design [candidate x (rater: station)] (pg. 10). It is a criterion-referenced examination (pgs. 8, 9).

2. Report more specifically the psychometric indices that are mentioned at the end of the first paragraph under the Procedures subheading.

This is now reported: e.g., Cronbach’s alpha > 0.70; item analysis – item difficulty and discrimination (pg. 8)

3. It is not clear if reliability of scores was assessed from performance across all 10 stations or only on the subset of 6 that were repeated at time 2.

This is clarified on pg. 10.

4. A number of clinical skill measures are mentioned. Exactly what measures were included for this study? Describing measures in their own subheading would be useful. Is
the examinee’s score the sum or percentage of correct checklist items across all 10 stations or just the 6 repeated stations?

This has been added on pg. 9. The sum of correct checklist items were across the 6 repeated stations.

5. Clarify if the standardized patients and physician rater is confounded; that is, their effects can or cannot be separated in the G-study.

This has been clarified on pages 9-10.

6. Describe the clinical case content for each of the 6 OSCEs with repeated material.

This has been included on pg. 9

7. Include IRT standard error of estimation for each station.

This has been included in Table 1.

8. Scale values are missing for both ordinate axes in Figure 1. Should these axes be labeled probability rather than proportion correct?

This figure has been deleted.

Minor Essential Revisions

1. The 2-PL and 3-PL models are presented in the paper, but there is no mention of the 1-PL Rasch model. The Multifacet Rasch model has been used in several studies to account for the effects of rater stringency, case difficulty and other facets typical to OSCE examinations. I would prefer to have seen some discussion about the consideration of these models for analyses.

The 1-PL Rasch model has been used in other studies, especially with MCQ items, to address several facets of tests. In the present study, we wished to include the second parameter (i.e., discrimination) in a 2-PL model as it may also be unstable over time in some complex way. Therefore, we wanted to be able to address two major psychometric characteristics (difficulty, discrimination) that is an advantage of the 2-PL model over the 1-PL Rasch model.

2. I am concerned that principal component factor analysis was compared between the two times of test administration for two of two stations without more discussion about these results. This analysis seems to be an attempt to test factor structural invariance between the two test administrations. Test of invariance over time is possible with confirmatory factor analysis. In this instance, the analysis is exploratory that could be investigated in the future with confirmatory factor analysis.
Upon closer reflection, we agree that EFA was not necessary and indeed only contributes confusion to the analyses. In the future, test of invariance over time should be done with confirmatory factor analysis. In the present study, we have removed the EFA from the paper (pgs. 11, 12-13, 20).

Response to the comments of reviewer two
Reviewer: Celia Taylor
Reviewer's report:

Major compulsory revisions

1. The introduction is somewhat repetitive and jumps around from one issue to another and therefore could be better structured and shortened.

The introduction has been tightened, better structured and shortened (see pgs. 3-7)

2. The introduction has not convinced me as to why stability over time is so critical – particularly when there are only two years’ worth of data. I would be more concerned about the performance of new stations. This partly depends on the approach to standard setting taken for this OSCE, e.g. the same Angoff used in both years vs. BLG/BLR undertaken each year – but details on standard setting are not provided.

Further discussion about the importance of stability is included on pages 5, 6 and 7. Further details of the stations including standard setting (i.e, Ebel) are discussed on pages 9 and 10.

3. Examiners only received 30 minutes of training – this does not seem very much and thus the sufficiency (or otherwise) of the training should be considered in the discussion.

We addressed this on page 15.

4. In an IRT model the mean item difficulty across all items (stations) should be 0; yet all stations have a negative item difficulty. This result needs to be explained.

In a 2PL model the guessing parameter is set to 0 (i.e., it is impossible to guess such as clinical skills), the discrimination parameters are bounded by 0 to 1 and the difficulty parameters are bounded at -3.0 and 3.0 (i.e., ±3 SDs around theta = 0). Therefore in the present study the negative difficulties indicate that the item characteristic curves are to the left of distribution as shown in Figure 2. Additionally, it is evident from Table 1 that station A and B are relatively easier than Stations C,D, and E which have difficulties closed to the mean of theta (0).

5. The authors conclude that non-significant results mean that the stations are stable but the results actually mean that a difference has not been detected. Without a power calculation it is not possible to determine the reliability of this interpretation.
The total sample size on which the MANOVA was based was \( n = 425 \) (189 + 236). As there were only 2 dependent variables (difficulty and discrimination simultaneously) in the analysis, the sample size easily exceeds that needed for adequate power to detect differences if they exist.

Minor essential revisions

1. Correction of typos and grammar errors and inconsistencies (e.g. 1980s vs. 1980’s; use of numerals for numbers <10).

   Done – thank you.

2. The introduction needs more clarity with respect to what is meant be ‘stability’.

   This has been done on pages 5, 6 and 7.

3. In the methods it is noted that “each year the AIMG program re-uses 4-6 OSCE stations from previous years with good psychometric indices”. The authors need to specify what these psychometric indices are and what makes them “good”.

   This is now reported: e.g., Cronbach’s alpha > 0.70; item analysis – item difficulty and discrimination (pg. 8)

4. The formulas need to be explained further, e.g. what is varcomp? Where is \( C_g \) in the IRT model formula? Is the IRT model used at checklist item level or at station level?

   \( \text{varcomp} = \text{variance component (corrected p. 10); } C_g \text{ is the guessing parameter typically in MCQ items but not included here because guessing is not a parameter in clinical skills assessment. For the 2PL model, } C_g \text{ (the lower asymptote for the guessing parameter) is set at zero. (pgs. 10, 11) } \)

5. The use of a simplified G model needs to be explained and its limitations discussed.

   While this type of design [single facet, nested: candidate x (rater: station)] does not allow for the estimation of the interaction effect of assessors and/or standardized patients with candidates, it does allow for the determination of the generalizability coefficient of assessors (p. 10)

6. The G coefficients are presented in the methods section but are actually results. What is meant by varying stations and why has the model not been used to evaluate overall generalizability? It is noted in the discussion that the G coefficients were “acceptable” yet the definition of “acceptable” is not defined in the paper.

   These have been moved to page 12 and acceptable has been defined (> .70).
7. In the methods section the authors need to explain why they have used MANOVA (e.g. what null hypothesis they are trying to test).

The null hypothesis in the MAOVAs is that the centroids (i.e., 2 dependent variables – difficulty and discrimination simultaneously) are not different at Time 1 and Time 2.

8. The use of exploratory factor analysis needs to be explained further. I had to assume that the TCCs computed with homogenous items used the results of the factor analysis. However I am not convinced what this approach (cf. using overall TCCs) adds to the paper. Again the sentence “The pattern of TCC for homogenous and heterogeneous items for all the OSCE stations was similar” belongs in the results section, not the methods section. If the factor analysis is retained, the full results should be shown in an Appendix. The third finding summarised in the discussion relates to the TCCs and presumably is based in a comparison of the TCCs in Figures 1 & 2 – this needs to be explained in the text. When this issue is considered again later in the discussion I think the word heterogeneous has been used instead of homogenous – without clarity here I struggled to understand the comments on clinical skills being contextual.

Upon closer reflection, we agree that EFA was not necessary and indeed only contributes confusion to the analyses. In the future, test of invariance over time should be done with confirmatory factor analysis. In the present study, we have removed the EFA from the paper (pgs. 11, 12-13, 20). We have deleted Figure 1.

9. In Table 1, explain what is meant by “Means” in the column headings for Time 1 and Time 2. The full test statistics and exact p-values for all tests should be provided. The authors need to specify how they have corrected their results for multiple comparisons.

A better description has been added in Table 1. The full test statistics are reported on page 12.

10. Figure 1: the y-axis needs labelling. Figure 2: x-axis scales have different centre points. It is a little confusing that the proportion correct = difficulty, since the term ‘difficulty’ has a specific meaning in IRT.

This figure has been deleted.

Discretionary revisions

1. The details of the Time 1 and Time 2 OSCEs are not very clear to a reader unfamiliar with the system and hence this section could be re-worded to improve readability.

This has been clarified on page 9.
2. Time 1 and Time 2 are used in the text but 2007 and 2008 on the figures. I suggest one or the other is used throughout.

3. I do not understand what is meant by “the slope of the curve” in the sentence “the overall discrimination (the slope of the tangent at the break in the curve” (sic) and this should be clarified.

   **This statement has been removed (pg. 12).**

4. Descriptions of the 6 stations would be helpful, but may not be provided for reasons of security.

   **This has been clarified on page 9 and Table 1.**

Thanks

Lubna Baig and Claudio Violato